

Lessons on Applying Automated Recommender Systems to Information-Seeking Tasks

Joseph A. Konstan^{*}, Sean M. McNee^{*}, Cai-Nicolas Ziegler[†],
Roberto Torres[‡], Nishikant Kapoor^{*}, and John T. Riedl^{*}

^{*}GroupLens Research, University of Minnesota, Minneapolis, MN 55455 USA

[†]Siemens AG, Corporate Technology, Otto-Hahn-Ring 6, D-81730 Munich, Germany

[‡]Novatec Editora, Rua Luis A. Santos, Sao Paulo – SP - Brazil

{konstan,mcnee,riedl,nkapoor}@cs.umn.edu; cai.ziegler@siemens.com; roberto.torres@novateceditora.com.br

Abstract^{*}

Automated recommender systems predict user preferences by applying machine learning techniques to data on products, users, and past user preferences for products. Such systems have become increasingly popular in entertainment and e-commerce domains, but have thus far had little success in information-seeking domains such as identifying published research of interest. We report on several recent publications that show how recommenders can be extended to more effectively address information-seeking tasks by expanding the focus from accurate prediction of user preferences to identifying a useful set of items to recommend in response to the user's specific information need. Specific research demonstrates the value of diversity in recommendation lists, shows how users value lists of recommendations as something different from the sum of the individual recommendations within, and presents an analytic model for customizing a recommender to match user information-seeking needs.

Background

For more than a decade, recommender systems researchers have applied machine learning techniques to predict user preferences for products ranging from movies to songs to jokes. These techniques have been adopted in a variety of commercial applications, including well-known recommenders such as at Amazon.com. Research in this area, however, has focused on improving prediction accuracy; researchers have mostly proposed new algorithms for extracting a small amount of increased accuracy out of an existing data set.

Over the past few years, we have been taking a different approach. In 2002 we published our first work on recommending research papers [McNee et al. 2002]. In that work, we demonstrated a system called TechLens that could produce useful recommendations for papers, though

we quickly discovered that novelty was as important as accuracy in providing value to users of the system.

In this Nectar paper, we review the past two years of progress on focusing automated recommender systems on producing recommendations of value to users, with a focus on users with information-seeking tasks. We then draw out lessons on the application of machine learning algorithms, and AI techniques in general, to serious user tasks.

Research Results

At the conclusion of the TechLens work, we knew three things:

- researchers (specifically Computer Science researchers) were eager enough to find interesting papers that they would value a system that produced even one recommendation out of five for a relevant paper they had not seen before;
- we could meet such a goal with traditional collaborative filtering algorithms, applied to a "ratings matrix" derived from paper citation graphs; and
- users had different reactions to the variety of algorithms tried (a mix of collaborative filtering, keyword techniques, and others)—in particular, user assessments of the relevance and novelty of recommendations varied substantially.

Exploring Hybrid Recommenders

In [Torres et al. 2004], we explored the use of hybrid recommender algorithms in an attempt to overcome the limitations of individual algorithms. In the process, we also used user experiments to evaluate the nature of recommendations produced by these algorithms (as seen in figure 1).

We found that pure collaborative filtering outperformed certain hybrids when evaluating recommendations for novelty and authoritativeness (with statistically significant results), but that a version started by augmenting the user's profile with a set of papers found based on term-set

^{*}This work has been supported by the National Science Foundation under grants 95-54517, 97-34442, and 01-02229.

Compilation copyright © 2006, American Association for Artificial Intelligence (www.aaai.org). All rights reserved.

overlap, and then applied traditional collaborative filtering to generate recommendations, tended to perform better at finding introductory, survey, or overview papers ($p < 0.1$). In this paper, we found two other interesting results:

- Brazilian research subjects, who were unfamiliar with a larger percentage of the papers recommended, were significantly happier with the recommendations.
- Satisfaction with sets of recommendations was different from (and higher than) satisfaction with individual recommendations.

Recommending Diverse Sets

Given the insight that recommendation sets might be viewed differently from the individual recommendations, we explored the question of diversifying recommendation sets by excluding from the set items too similar to those already recommended. Specifically, in [Ziegler et al. 2005] we explored the use of taxonomic data to diversify top- n recommendation lists. In off-line analyses, we showed that diversifying in this manner would reduce the intra-list similarity of the recommendation set, but would also decrease precision and recall.

We experimented with both user-user correlation and item-item correlation recommenders. The user-user algorithm showed no measurable benefit from increased diversification, and indeed started out with higher user satisfaction scores for the recommendation list. The item-item algorithm, however, when given a 40% diversification factor, improved its user scores not only above undiversified, but also above the user-user algorithm's score. Analysis confirmed that the diversified lists yielded higher satisfaction scores for the list as a whole, even through the user ratings of individual items were lower. These higher-scoring lists covered a broader range of the user's interests.

Figure 1. User evaluation of recommended papers.

From this study, we take away the clear message that understanding a user's information need is critical. Most users of a book recommender were looking for a variety of books to consider—the same diversification approach would likely fail if the user were a professor seeking recommendations of course textbooks—in that case low diversity and high similarity might well be considered good properties.

Human-Recommender Interaction and Matching Recommenders to User Tasks

The current state of our work is described in [McNee et al. 2006a] which presents an analytic model for human-recommender interaction. This model, illustrated in Figure 2, creates a multi-step mapping from users and their tasks to the appropriate recommender algorithms (and application tuning factors) to serve those users.

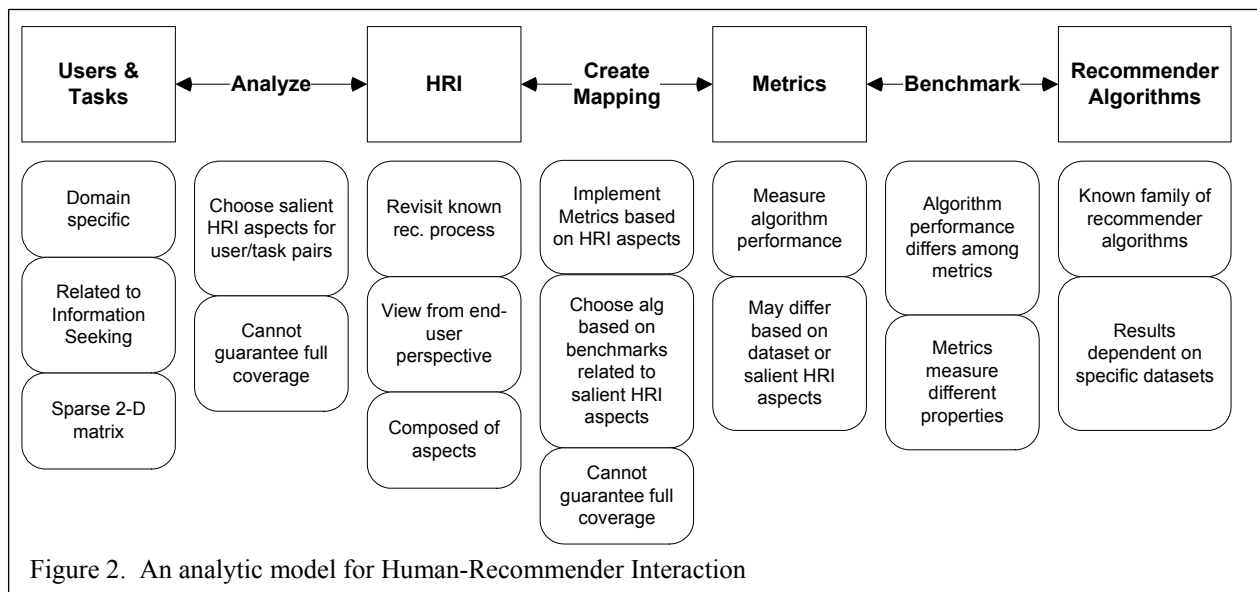


Figure 2. An analytic model for Human-Recommender Interaction

There are four key elements to this model:

- The modeling of user tasks in a domain. In our case, we use research on information-seeking from the library and information science literature [Case, 2002; Kuhlthau, 2004] to model the types of tasks users present.
- The modeling of recommender system aspects as seen from a user perspective. HRI divides these into dialogue and personality attributes. Dialogue attributes model the properties of a particular set of returned recommendations, and include such features as correctness, serendipity, usefulness, and spread. Personality attributes model properties of repeated interaction, and include such features as personalization, risk-taking, adaptability, and pigeonholing.
- The definition of a set of metrics to measure algorithm performance against goals beyond simply accuracy. Following the lead of [Herlocker et al. 2004], we argue that accuracy alone is an inadequate measure of recommender system performance [McNee et al. 2006b].
- The mappings among these categories. For our domain, we have established analytic mappings between users/tasks and attributes, and benchmark studies to map recommender algorithms to metrics. We are now in the process of a study (still unpublished) that uses user studies to validate the user/task-to-attribute mapping, and to establish an initial metric-to-attribute mapping.

HRI provides a framework for recommender system design based on knowledge of users and their tasks. At present, the HRI framework requires considerable analysis and experimentation to understand the relationships among tasks, data sets, and algorithms. Over time, we hope this work is extended to identify properties of domains and datasets that will allow greater use of analytic tools and less experimentation in design.

Lessons for the AI Community

The core lesson of the above research concerns the application of user-centered design techniques to AI systems. We consider three aspects of this lesson in detail: the appropriate evaluation of recommender system algorithms, the need for richer collections of test data, and the opportunity to develop innovative algorithms to address specific information-seeking needs.

Evaluation and Metrics

Recommender systems, from one point of view, are an application of machine learning systems to predict future user behavior based on a matrix of past behaviors. The most common evaluation of the effectiveness of such systems has been to assess the accuracy with which they can estimate withheld data (the leave-*n*-out approach).

Yet, from the user's point of view, this metric assesses the least useful property of a recommender—its ability to "recommend" the items the user already has experienced and knows. (Yes, as researchers, we pretend that withholding an item is the same as finding an item the person hasn't rated, but we know it isn't the same.) This is not just a pedantic distinction—when finding research papers, a user cares a great deal about the difference between receiving recommendations for papers she is already aware of and those that she is unaware of. And the leave-*n*-out methodology weights evaluation towards the type of papers she already knows, rather than the ones that would be new to her.

Of course, we're well aware that there is not an easy solution to this challenge. The reason we use the leave-*n*-out method is precisely to avoid the problem of having recommendations for which we cannot assess the value to the user. A few experiments have been able to directly assess user opinion of novel recommendations. The joke recommender Jester [Goldberg et al. 2001] was able to do so because the recommended items (jokes) were rapidly consumable. We have also conducted limited experiments using movie recommendation (in which we paid for the subject to watch the movie). But in general, such experimentation is difficult.

The difficulty of such direct utility measures is part of the reason for using a set of different evaluation metrics—metrics that explore a variety of features that may be relevant to users. Our work with HRI suggests some of these metrics: serendipity, authority, boldness, coverage, adaptivity, and personalization.

Test Data Collections

Datasets can substantially shape the progress of a field. Many would claim that the machine learning repository at UC Irvine has helped advance machine learning research by giving researchers sets of data against which to compare their ideas and algorithms. Similarly, the availability of a few key recommender systems datasets (originally, the EachMovie dataset, now the MovieLens, Jester, and BookCrossing datasets) has advanced the field by removing the original barrier to entry—the need to develop a system and user base before experimenting with algorithms and analytic tools.

At the same time, the limitations of the datasets also shape the field. Some have argued that the early TREC competitions and datasets, while a boon to information retrieval generally, hurt the ability of researchers to advance personalized retrieval solutions (e.g., recommender systems) because those datasets lacked individual relevance (i.e., rating) data. One can similarly argue that today's available recommender system datasets are a limiting factor in recommender systems research. None of today's available datasets include any information on user tasks, and all of them are built on the assumption that all users are addressing the same task across all their interactions. Furthermore, none of today's recommender systems datasets address non-entertainment domains.

As a result of these factors, researchers on recommender systems, and other related areas, should recognize the limitations imposed by working with existing datasets (and, indeed, probably need to do more work with human subjects and new data sets). Furthermore, there is a substantial opportunity for creation and dissemination of new and different datasets.

The Opportunity for New Algorithms

As we have been exploring the domain of research-paper recommenders, we have been working with collection managers and librarians to understand the breadth of tasks that users bring to these collections. While our initial work has focused on variations of the "find me more like this" task—a task in which a user has a starting point of a paper or a bibliography, and seeks other papers to read—we have identified a large set of interesting tasks where tools built using AI techniques. For example:

Library collections can recommend people and even locations as well as individual papers. What algorithms provide effective answers to questions such as "where would be a good place to spend a sabbatical (or postdoc) based on the work I've done, and a few examples of the kind of work I admire?" or even "who would be good members for the program committee of this conference, given the desire for topical and geographic diversity, and using the past three years of proceedings as an example of the type of content in the conference?"

Recommendation may also be too narrow an output; in order to fulfill user needs, we may need to be able to explain or justify the recommendations (see, for example [Herlocker et al. 2000]). How do we explain in meaningful terms the reasons for recommending a paper, program committee member, or department? Indeed, how do we explain why we didn't recommend something that the user expected we'd recommend?

Finally, we see a long-term tie between recommender systems and other intelligent systems. We imagine a user asking their recommender-laden library to be able to ask questions such as "what do I need to know about Bayesian Belief Networks research?" and get back pointers to good overview and survey articles, a list of the "big names" in the field, pointers to the conferences or other venues where the important work in the field is published, etc. And of course, this should work even for a field that isn't mature enough to appear in the formal keyword systems and taxonomies.

Conclusion

Recommender systems help people find items of interest from large information spaces. This direct interaction with end users creates new and difficult challenges than has been explored by machine learning and AI research in the past. These difficulties become more serious as recommenders move into information spaces where users may bring a wide variety of information needs. We claim that by understanding the user's information seeking task,

we can generate a more useful recommendation list. To do this, not only do we need to understand user tasks, we need to rethink about how we evaluate recommender algorithms, possibly using a wider variety of metrics over multiple datasets, and we need to integrate new and more flexible algorithms into recommender systems so that we can select the most appropriate algorithm for a user's need.

References

- D.O. Case (2002) *Looking for Information: A Survey of Research on Information Seeking, Needs, and Behavior*, San Diego: Academic Press, 2002.
- K. Goldberg, T. Roeder, D. Gupta, and C. Perkins (2001) Eigentaste: A Constant Time Collaborative Filtering Algorithm *Information Retrieval Journal*, 4(2), pp. 133-151. July 2001.
- J. Herlocker, J. Konstan, and J. Riedl (2000) Explaining Collaborative Filtering Recommendations. In proceedings of *ACM 2000 Conference on Computer Supported Cooperative Work*, December 2-6, 2000, pp. 241-250.
- J. Herlocker, J. Konstan, L. Terveen and J. Riedl (2004). Evaluating Collaborative Filtering Recommender Systems. *ACM Transactions on Information Systems* 22(1), pp. 5-53, January 2004.
- J.A. Konstan, N. Kapoor, S.M. McNee, and J.T. Butler (2005). TechLens: Exploring the Use of Recommenders to Support Users of Digital Libraries. *CNI Fall Task Force Meeting Project Briefing*. Coalition for Networked Information. Phoenix, AZ.
- C.C. Kuhlthau (2004) *Seeking Meaning: A Process Approach to Library and Information Services*, Westport, CT: Libraries Unlimited, 2004.
- S.M. McNee, I. Albert, D. Cosley, P. Gopalkrishnan, S.K. Lam, A.M. Rashid, J.A. Konstan, and J. Riedl (2002). On the Recommending of Citations for Research Papers. In *Proceedings of ACM 2002 Conference on Computer Supported Cooperative Work (CSCW2002)*, New Orleans, LA, pp. 116-125.
- S.M. McNee, J. Riedl, and J.A. Konstan (2006a). "Making Recommendations Better: An Analytic Model for Human-Recommender Interaction". In the Extended Abstracts of *the 2006 ACM Conference on Human Factors in Computing Systems (CHI 2006)*, April 2006.
- S.M. McNee, J. Riedl, and J.A. Konstan (2006b). "Being Accurate is Not Enough: How Accuracy Metrics have hurt Recommender Systems". In the Extended Abstracts of *the 2006 ACM Conference on Human Factors in Computing Systems (CHI 2006)*, April 2006.
- R. Torres, S.M. McNee, M. Abel, J.A. Konstan, and J. Riedl (2004). Enhancing Digital Libraries with TechLens+. In Proceedings of *The Fourth ACM/IEEE Joint Conference on Digital Libraries (JCDL 2004)*, June 2004, pp. 228-237.
- C-N Ziegler, S.M. McNee, J.A. Konstan, and G. Lausen (2005). "Improving Recommendation Lists Through Topic Diversification." In Proceedings of *the Fourteenth International World Wide Web Conference (WWW2005)*, May 2005, pp. 22-32.